



Scholars' Mine

Masters Theses

Student Theses and Dissertations

1970

Z-form solution of non-linear differential equations

Allen Joseph Rushing

Follow this and additional works at: https://scholarsmine.mst.edu/masters_theses

 Part of the [Electrical and Computer Engineering Commons](#)

Department:

Recommended Citation

Rushing, Allen Joseph, "Z-form solution of non-linear differential equations" (1970). *Masters Theses*. 7060.

https://scholarsmine.mst.edu/masters_theses/7060

This thesis is brought to you by Scholars' Mine, a service of the Missouri S&T Library and Learning Resources. This work is protected by U. S. Copyright Law. Unauthorized use including reproduction for redistribution requires the permission of the copyright holder. For more information, please contact scholarsmine@mst.edu.

Z-FORM SOLUTION OF
NON-LINEAR DIFFERENTIAL EQUATIONS

BY

ALLEN JOSEPH RUSHING, 1944-

A

THESIS

submitted to the faculty of

UNIVERSITY OF MISSOURI - ROLLA

in partial fulfillment of the requirements for

Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

Rolla, Missouri

1970

Approved by

John A. Knull (advisor) Bonny Hachant
Herbert A. Crosby

Abstract

The "Z-Forms" developed by Boxer and Thaler from Z-Transform theory are reviewed. Iteration is shown to increase accuracy when Z-Forms are used to obtain numerical solutions to some non-linear differential equations. Machine results are presented for example problems. Error estimates and convergence conditions are discussed. The method is usable only when 2- or 3-digit accuracy is acceptable. For some classes of equations the Z-Form method compares favorably to other numerical methods in time required to set up the problem and in machine time required for solution.

Preface

In the past 25 years electrical engineers have led in the development of discrete-time system theory. The wide use of digital computers, both as tools for analysis and simulation, and as real-time system building blocks has made the study of discrete-time systems more important than ever before.

In the 1950's several methods were developed for converting continuous-time (Laplace-Transform) transfer functions into equivalent discrete-time (Z-Transform) transfer functions. The resulting transfer function is a ratio of polynomials in Z or Z^{-1} . The inverse Z-Transform is obtained simply by carrying out the indicated long division. The discrete-time solution is given by the coefficients of the quotient terms.

Search of the literature revealed surprisingly little work on the subject despite the apparent usefulness of the method. From the literature it is unclear why the method was not refined and placed alongside the Runge-Kutta and predictor-corrector methods as a tool of numerical analysts for the solution of differential equations. Flaws limiting the application of the method have been briefly pointed out but inadequately described in the literature.

Part of the reason for this lack of literature may

be the difficulty of communication between technical disciplines. Theoretical background is sparse because the method was developed by electrical engineers, who were not mainly interested in mathematical refinement, while mathematicians and numerical analysts remained largely unaware of the engineers' work. Applications of the method were never made by workers outside the field of electrical engineering, so far as the author is aware.

In 1955 Boxer and Thaler made an original contribution to the conversion of continuous systems into discrete-time systems when they introduced their "Z-Forms". Boxer and Thaler also posed several questions needing study before the usefulness of their Z-Forms could be judged soundly. One of these questions initiated this thesis research.

When their method is applied to non-linear differential equations, Boxer and Thaler predicted that iteration of each stage of the long division process would improve the accuracy of the solution. They guessed, however, that to achieve the same accuracy, it would be more efficient to decrease the sampling period than to iterate.

The most important result of the research for this thesis is that iteration of each stage of the long division is often an efficient way to increase accuracy.

There is a component of error, referred to as recursion error in this paper, which can be reduced most efficiently by iteration, rather than by a reduction of the sampling period.

The author is grateful to the Monsanto Company for permitting the use of the remote computer terminal at their Page Technical Center. As an employee at the Page Technical Center the author found the computer terminal there convenient for this thesis research.

TABLE OF CONTENTS

	Page
ABSTRACT	ii
PREFACE	iii
LIST OF FIGURES.	viii
LIST OF TABLES	ix
I. INTRODUCTION	1
A. THEORY	1
B. EXAMPLES	6
II. REVIEW OF LITERATURE	16
A. FORMS SIMILAR TO THE Z-FORMS OF BOXER AND THALER	16
B. ITERATION.	19
C. NEW METHODS FOR SYSTEM SIMULATION.	20
D. PROGRAMMING Z-TRANSFORM INVERSION.	21
III. DISCUSSION	22
A. AN ALTERNATIVE DERIVATION OF THE Z-FORMS, TAKING INTO ACCOUNT NON-ZERO INITIAL CONDITIONS	22
B. ERROR.	31
1. DISCRETIZATION.	31
2. RECURSION	34
3. CONVERGENCE OF ITERATION.	37
C. OTHER DIFFICULTIES AND LIMITATIONS OF ITERATION.	41
D. CHOICE OF SAMPLING PERIOD.	44
E. NOTES ON THE COMPUTER PROGRAM.	46
IV. CONCLUSIONS.	48

TABLE OF CONTENTS (continued)

	Page
V. APPENDIX -- LISTING OF THE COMPUTER PROGRAM . .	51
VI. BIBLIOGRAPHY. ,	55
VII. VITA.	57

LIST OF FIGURES

1. Error in the solution of $\ddot{y} + \ddot{y} + \dot{y} = 1$,
for three families of Z-Forms. 18
2. Solution of Van der Pol's equation 43
3. Accuracy of trapezoidal rule integration 45

LIST OF TABLES

	Page
I. Z-Forms of Boxer and Thaler, Madwed and Truxal, and Tustin	3
II. Machine results for $\dot{y}+2y=1$, $y(0)=0$	8
III. Machine results for $\dot{y}+y^2=1$, $y(0)=0$	11
IV. Machine results for $\ddot{y} + \dot{y} + y = 1$, $\dot{y}(0)=\ddot{y}(0)=y(0)=0$ for three families of Z-Forms	17
V. Machine results for $\ddot{y}+y=0$, $y(0)=0$ $\dot{y}(0)=1$	28
VI. Machine results for $\dot{y}+2y=1$, $y(0)=0$, by both original and modified methods	30
VII. Expansion of the Boxer and Thaler Z-Forms about $s=0$	33
VIII. Machine results for $\dot{y}=e^t$, $y(0)=1$	35
IX. Machine results for $y^2\dot{y}+y=0$, $y(0)=1$ $\dot{y}(0)=0$	40
X. Machine results for $\ddot{y}+0.1\dot{y}^2=32.2$, $y(0)=\dot{y}(0)=0$	42
XI. Numbers of arithmetic operations required by different methods in various problems	49

I. Introduction

A. Theory

The variable of the Z-transform, z , is related to the variable of the Laplace transform, s , by the following defining equation.

$$z = e^{sT} \quad (1)$$

where T is the time interval from one instant in discrete time to the next. One can use the above equation to convert continuous-time Laplace-transformed functions in s into equivalent discrete-time Z-transformed functions in z , and vice versa.

After converting from continuous to discrete transfer functions difficulty arises when one attempts to find the inverse Z-transform to obtain the discrete-time solution. At this point one usually has a ratio of polynomials containing powers of $(\frac{1}{T} \ln z)$. There is no straight-forward way of obtaining the inverse Z-transform of such a transcendental function. Theoretically one could use the inversion integral for the Z-transform, but this would be impractical. One would have to repeatedly evaluate the integral of a transcendental function of a complex variable around a contour in the complex plane [1]. The following example will illustrate.

$$\text{Suppose } F(s) = \frac{1}{s}.$$

$$\text{Then applying (1), } F(z) = \frac{T}{\ln z}.$$

The inverse Z-Transform is given by $f(k) = \frac{1}{2\pi j} \oint_{\Gamma} z^{k-1} F(z) dz$ where the contour is taken around a circle centered at the origin in the z-plane and including all the singularities of $F(z)$. Then by substitution we obtain

$$f(k) = \frac{1}{2\pi j} \oint_{\Gamma} z^{k-1} \frac{T}{\ln z} dz$$

To obtain the discrete-time solution $f(k)$ this integral must be evaluated for each instant of discrete-time.

The Z-Form method is derived from Z-Transform theory. The method uses a truncated infinite series approximation to $\frac{1}{T} \ln z$ to obtain Z-Transform functions which are ratios of polynomials in powers of z or z^{-1} . After performing the indicated long division, taking the inverse Z-Transform is trivial [2].

$$\text{From (1), } s = \frac{1}{T} \ln z$$

$$\text{Let } z = \frac{1+u}{1-u}, \text{ or } u = \frac{z-1}{z+1}$$

$$\text{Then } s = \frac{1}{T} \ln \frac{1+u}{1-u}$$

Using a Laurent Series approximation,

$$s = \frac{T}{2} \left(u + \frac{1}{3}u^3 + \frac{1}{5}u^5 + \frac{1}{7}u^7 + \dots \right)$$

$$s^{-1} = \frac{T}{2} \left(\frac{1}{u} - \frac{1}{3}u - \frac{4}{45}u^3 - \frac{44}{945}u^5 + \dots \right) \quad (2)$$

If $|u| \ll 1$, the series may be truncated after the first term yielding

$$s^{-1} \approx \frac{T}{2u}$$

$$s^{-1} \approx \frac{T(z+1)}{2(z-1)}$$

which is the first Boxer and Thaler Z-Form of Table 1.

It would be well at this point to justify the

Table 1

Z-Forms of Boxer and Thaler, Madwed and Truxal, and Tustin.

$\frac{1}{s^n}$	Boxer & Thaler	Madwed- Truxal	Tustin
$\frac{1}{s}$	$\frac{T(z+1)}{2(z-1)}$	$\frac{T(z+1)}{2(z-1)}$	$\frac{T(z+1)}{2(z-1)}$
$\frac{1}{s^2}$	$\frac{T^2(z^2+10z+1)}{12(z-1)^2}$	$\frac{T^2(z^2+4z+1)}{6(z-1)^2}$	$\frac{T^2(z+1)^2}{4(z-1)^2}$
$\frac{1}{s^3}$	$\frac{T^3 z(z+1)}{2(z-1)^3}$	$\frac{T^3(z^3+11z^2+11z+1)}{24(z-1)^3}$	$\frac{T^3(z+1)^3}{8(z-1)^3}$
$\frac{1}{s^4}$	$\frac{T^4 z(z^2+4z+1)}{6(z-1)^4}$	$-\frac{T^4}{720} \frac{T^4(z^4+26z^3+66z^2+26z+1)}{120(z-1)^4}$	$\frac{T^4(z+1)^4}{16(z-1)^4}$

assumption that $|u| \ll 1$. The variable u was defined,

$$u = \frac{z-1}{z+1} = \frac{e^{sT}-1}{e^{sT}+1}$$

If the product sT is sufficiently small in magnitude the numerator approaches zero and the denominator approaches 2, so that $|u| \ll 1$. The quantity sT is complex.

$$s = \sigma + j\omega$$

where σ is exponential rate of increase in nepers/sec.

and ω is the sinusoidal frequency in radians/sec.

$$u = \frac{e^{\sigma T} e^{j\omega T} - 1}{e^{\sigma T} e^{j\omega T} + 1}$$

In the case of pure sinusoids, (and hence, by Fourier analysis, any waveform which satisfies the Dirichlet conditions for well-behaved periodic functions) $\sigma=0$ and

$$u = \frac{e^{j\omega T} - 1}{e^{j\omega T} + 1}$$

The magnitude of the variable u will be specified to be at most 0.1, so as to limit the error of the first-order Z-Form to 0.3%.

$$0.1 \geq \frac{e^{j\omega T} - 1}{e^{j\omega T} + 1} \approx \frac{1 + j\sin\omega T - 1}{1 + j\sin\omega T + 1} = \frac{j\sin\omega T}{2 + j\sin\omega T}$$

$$0.9j\sin\omega T < 0.2$$

$$T < \frac{0.22}{j\omega}$$

This result means that the sampling period must be much smaller than the period of the highest frequency component of interest, if the Z-Forms are to be valid. It means that for a Z-Form of s^{-1} to have 0.3% error, the sampling frequency should be about 28.5 times as great as the highest

frequency component of interest in the problem.

The Z-Form of s^{-2} will now be derived. From (2),

$$s^{-1} = \frac{T}{2} \left[\frac{1}{u} - \frac{1}{3}u - \frac{4}{45}u^3 - \frac{44}{945}u^5 + \dots \right]$$

Squaring both sides,

$$s^{-2} = \frac{T^2}{4} \left[\frac{1}{u^2} - \frac{2}{3} - \frac{1}{15}u^2 + \frac{8}{135}u^4 + \dots \right] \quad (3)$$

Truncating after the first two terms of the series,

$$\begin{aligned} s^{-2} &\approx \frac{T^2}{4} \left[\frac{1}{u^2} - \frac{2}{3} \right] = \frac{T^2}{4} \left[\frac{(z+1)^2}{(z-1)^2} - \frac{2}{3} \right] \\ &= \frac{T^2}{12} \left[\frac{z^2 + 10z + 1}{(z-1)^2} \right] \end{aligned}$$

which is the Boxer and Thaler Z-Form for s^{-2} in Table I.

The other Boxer and Thaler Z-Forms in Table I are derived similarly, retaining the principal part and the constant term of the series in each case. Boxer and Thaler [2] demonstrated that the terms beyond the constant terms in (2) and (3) and in the corresponding equations for higher-order Z-Forms must be truncated in order to avoid erroneous oscillation in the solution.

Higher-order Z-Forms require fewer samples per cycle than the Z-Forms for s^{-1} , to achieve the same accuracy. This is evident when one compares equations (2) and (3).

It is noteworthy that the Z-Forms have the same phase shift as the corresponding Laplace operators for $s=j\omega$.

The first-order Z-Form may be written as

$$\frac{1}{s} \approx \frac{T(e^{sT} + 1)}{2(e^{sT} - 1)} = \frac{T(\cos\omega T + j\sin\omega T + 1)}{2(\cos\omega T + j\sin\omega T - 1)}$$

Algebraic manipulation yields

$$\frac{1}{s} \approx \frac{-jT\sin\omega T}{2(1 - \cos\omega T)}$$

where the $-j$ factor in the numerator indicates a 90-degree phase lag, which is the ideal phase shift for integrating a sinusoidal waveform. A similar derivation shows the second-order Z-Form to have a phase shift of 180 degrees.

B. Examples

Simple examples will be worked out to show how the Z-Forms are used to obtain solutions differential equations. In the following examples the initial conditions are assumed to be zero. It will become apparent later that the Z-Forms require modification when applied to problems with non-zero initial conditions.

$$\dot{y} + 2y = 1, \quad y(0) = 0$$

Laplace transforming,

$$sY(s) - y(0) + 2Y(s) = \frac{1}{s}$$

$$Y(s) = \frac{1}{s(s+2)} = \frac{s^{-2}}{1+2s^{-1}}$$

Substituting the Z-Forms from Table I,

$$TY(z) = \frac{\frac{T^2}{12} \frac{z^2+10z+1}{(z-1)^2}}{1 + T \frac{z+1}{z-1}}$$

The substitution of the Z-Forms is effectively an integrating operation of $Y(z)$ over the time interval T . Hence on the left-hand side of the equation $Y(z)$ is multiplied by T [3].

Continuing,

$$Y(z) = \frac{\frac{T}{12}(z^2 + 10z + 1)}{(z-1)^2 + T(z+1)(z-1)}$$

$$= \frac{\frac{T}{12}z^2 + \frac{10T}{12}z + \frac{T}{12}}{[1+T]z^2 - 2z + [1-T]}$$

Choosing $T=0.1$ and performing the indicated division with slide rule accuracy,

$$\begin{array}{r}
 1.1z^2 - 2z + .9 \overline{) \begin{array}{l} .00757 + .089z^{-1} + .164z^{-2} + \dots \\ .00833z^2 + .0833z + .00833 \\ \hline .00833z^2 + .0151z + .00682 \\ \hline .0984z + .00151 \\ .0984z - .1788 + .0805z^{-1} \\ \hline .1803 - .0805z^{-1} \\ .1803 - .328z^{-1} \dots \\ \hline \dots \end{array}}
 \end{array}$$

The factor z^{-1} is a delay operator (delay by T). Hence the coefficients of the quotient terms comprise the discrete-time solution of the differential equation. That is, the first coefficient is the solution at time equal to zero, the second coefficient the solution at time equal to T , the third coefficient the solution at time equal to $2T$, etc. Table II presents the machine results for the preceding problems, along with the exact solution, obtained by evaluating the analytic solution, $y(t) = 0.5(1 - e^{-2t})$

Non-linear differential equations may be solved by linearizing about successive points [2]. An example follows.

$$\dot{y} + y^2 = 1, \quad y(0) = 0$$

The equation is approximately linear in a region over which $y(t)$ does not change very much. If the value of $y(t)$ is C

Table II

Machine results for $\dot{y}+2y=1$, $y(0)=0$.

Time	Exact Solution	Z-Form Solution, T=.1
.000	.000000	.007576
.100	.090635	.089532
.200	.164840	.164162
.300	.225594	.225224
.400	.275336	.275183
.500	.316060	.316059
.600	.349403	.349503
.700	.376702	.376866
.800	.399052	.399254
.900	.417351	.417571
1.000	.432332	.432558
1.100	.444598	.444820
1.200	.454641	.454853
1.300	.462863	.463062
1.400	.469595	.469778
1.500	.475106	.475273
1.600	.479619	.479769
1.700	.483313	.483447
1.800	.486338	.486457
1.900	.488815	.488919
2.000	.490842	.490934
2.100	.492502	.492582
2.200	.493861	.493931
2.300	.494974	.495034
2.400	.495885	.495937
2.500	.496631	.496676
2.600	.497242	.479280
2.700	.497742	.497775
2.800	.498151	.498179
2.900	.498486	.498510
3.000	.498761	.498781
3.100	.498985	.499003
3.200	.499169	.499184
3.300	.499320	.499332
3.400	.499443	.499454
3.500	.499544	.499454
3.600	.499627	.499553
3.700	.499694	.499701
3.800	.499750	.499755
3.900	.499795	.499800
4.000	.499832	.499836

at some point in the region, the equation can be linearized about this point by writing

$$\dot{y} + Cy = 1$$

If C is treated as a constant the Laplace transform can be applied.

$$sY(s) - y(0) + CY(s) = \frac{1}{s}$$

$$Y(s) = \frac{1}{s(s+C)} = \frac{s^{-2}}{1+Cs^{-1}}$$

Now the Z-Forms can be substituted as in the previous example, yielding after some algebraic manipulation

$$Y(z) = \frac{T(z^2 + 10z + 1)}{(12 + 6CT)z^2 - 24z + (12 - 6CT)}$$

In performing each stage of the long division the value of C is taken to be the last previously calculated value of y . For the first stage of the division C is taken to be $y(0)=0.0$.

The accuracy of the Z-Form solution to this equation can be improved by iterating each stage of the long division process. This iteration involves calculating an initial approximate solution at a point in discrete-time, as before. This approximation is then substituted for C into the denominator coefficients, and the division repeated to obtain an improved approximate solution at the same point in discrete-time. The iteration may be repeated until some convergence criterion is satisfied, or until some maximum number of iterations have been performed,

whichever comes first. Then one goes on to the next stage of the long division to obtain the solution at the next point in discrete-time.

Table III presents the machine results for the Z-Form solution of the equation $\dot{y} + y^2 = 1$, $y(0)=0$ for various values of T and various numbers of iterations. The exact solution, obtained by evaluating the analytic solution, $y(t)=\tanh(t)$, is also presented.

Table III

Machine results for $\dot{y}+y^2=1$, $y(0)=0$ Z-Form Solutions, $T=.2$

Time	Exact Solution	No Iteration	One Iteration	Multiple Iterations
.000	.000000	.016667	.016667	.016667
.200	.197375	.199667	.196085	.196152
.400	.379949	.391517	.377644	.378015
.600	.537050	.561708	.534136	.535093
.800	.664037	.700376	.660911	.662562
1.000	.761594	.804677	.758561	.760784
1.200	.833655	.877694	.830908	.833442
1.400	.885352	.925810	.882989	.885558
1.600	.921669	.956041	.919714	.922108
1.800	.946806	.974375	.945239	.947336
2.000	.964028	.985223	.962801	.964554

Notes: In the solutions obtained by multiple iteration, iteration was terminated with a relative error criterion of .000001, up to a maximum of 9 iterations.

Iteration was not performed for the first point in discrete time (time=0.0 in this problem) because the exact initial value was known, i.e., $y(0)=0$, and this value was used in the initial calculation. This accounts for the identical numbers at the top of the last three columns.

continued

Table III (continued)

Z-Form Solutions, T=.1

Time	Exact Solution	No Iteration	One Iteration	Multiple Iterations
.000	.000000	.008333	.008333	.008333
.100	.099668	.099958	.099503	.099505
.200	.197375	.198923	.197054	.197068
.300	.291313	.294994	.290854	.290895
.400	.379949	.386361	.379376	.379465
.500	.462117	.471553	.461456	.461611
.600	.537050	.549488	.536329	.536562
.700	.604368	.619512	.603613	.603930
.800	.664037	.681385	.663272	.663671
.900	.716298	.735230	.715543	.716014
1.000	.761594	.781454	.760864	.761394
1.100	.800499	.820661	.799806	.800378
1.200	.833655	.853571	.833006	.833603
1.300	.861723	.880949	.861125	.861729
1.400	.885352	.903552	.884805	.885403
1.500	.905148	.922095	.904654	.905234
1.600	.921669	.937226	.921226	.921778
1.700	.935409	.949520	.935015	.935533
1.800	.946806	.959472	.946457	.946937
1.900	.956237	.967506	.955931	.956371
2.000	.964028	.973975	.963759	.964158

continued

Table III (continued)

Z-Form Solutions $T=.05$

Time	Exact Solution	No Iteration	One Iteration	Multiple Iterations
.000	.000000	.004167	.004167	.004167
.100	.099668	.099865	.099627	.099627
.200	.197375	.198254	.197295	.197298
.300	.291313	.293266	.291198	.291208
.400	.379949	.383242	.379806	.379828
.500	.462117	.466862	.461953	.461991
.600	.537050	.543208	.536871	.536928
.700	.604368	.611778	.604181	.604258
.800	.664037	.672447	.663848	.663945
.900	.716298	.725413	.716112	.716227
1.000	.761594	.771107	.761415	.761544
1.100	.800499	.810124	.800330	.800469
1.200	.833655	.843144	.833497	.833642
1.300	.861723	.870877	.861579	.861725
1.400	.885352	.894022	.885220	.885365
1.500	.905148	.913232	.905030	.905170
1.600	.921669	.929106	.921563	.921696
1.700	.935409	.942175	.935316	.935440
1.800	.946806	.952900	.946724	.946839
1.900	.956237	.961680	.956165	.956271
2.000	.964028	.968852	.963965	.964060

continued

Table III (continued)

Z-Form Solutions, $T=.02$

Time	Exact Solution	No Iteration	One Iteration
.000	.000000	.001667	.001667
.100	.099668	.099760	.099661
.200	.197375	.197746	.197363
.300	.291313	.292114	.291294
.400	.379949	.381278	.379926
.500	.462117	.464012	.462091
.600	.537050	.539490	.537021
.700	.604368	.607286	.604338
.800	.664037	.667334	.664007
.900	.716298	.719859	.716268
1.000	.761594	.765302	.761566
1.100	.800499	.804244	.800472
1.200	.833655	.837344	.833630
1.300	.861723	.865282	.861700
1.400	.885352	.888724	.885331
1.500	.905148	.908295	.905130
1.600	.921699	.924567	.921652
1.700	.935409	.938050	.935395
1.800	.946806	.949189	.946793
1.900	.956237	.958370	.956226
2.000	.964028	.965922	.964018

continued

Table III (continued)

Z-Form Solution, $T=.01$

Time	Exact Solution	No Iteration	One Iteration
.000	.000000	.000833	.000833
.100	.099668	.099716	.099666
.200	.197375	.197564	.197372
.300	.291313	.291716	.291308
.400	.379949	.380615	.379943
.500	.462117	.463063	.462111
.600	.537050	.538265	.537042
.700	.604368	.605818	.604360
.800	.664037	.665673	.664029
.900	.716298	.718064	.716290
1.000	.761594	.763431	.761587
1.100	.800499	.802354	.800492
1.200	.833655	.835482	.833648
1.300	.861723	.863485	.861717
1.400	.885352	.887022	.885346
1.500	.905148	.906707	.905143
1.600	.921669	.923105	.921664
1.700	.935409	.936718	.935405
1.800	.946806	.947988	.946802
1.900	.956237	.957296	.956234
2.000	.964028	.964969	.964025

II. Review of Literature

A. Forms similar to the Z-Forms of Boxer and Thaler

It was shown in the Introduction how Boxer and Thaler originally derived their Z-Forms. Different sets of Z-Forms were derived earlier by Tustin and by Madwed and Truxal [4]. These sets of Z-Forms are identical to the Boxer and Thaler Z-Forms for the first-order integration operator (s^{-1}), but different for higher-order operators.

The Boxer and Thaler Z-Forms generally lead to the most accurate solutions; the Madwed-Truxal forms lead to intermediate accuracy; the Tustin forms lead to the least accurate solutions. Table IV presents the results to the solution of the equation $\ddot{y} + \ddot{y} + \dot{y} = 1$, $\ddot{y}(0) = \dot{y}(0) = y(0) = 0$ using each of the three families of Z-Forms, along with the exact solution,

$$y(t) = 1 - e^{-.5t} \cos \sqrt{.75} t - \frac{.5}{\sqrt{.75}} e^{-.5t} \sin \sqrt{.75} t$$

The error of each of the three solutions is plotted in Figure 1. Note that at some points in discrete-time the Boxer and Thaler results are not the most accurate, but that over a span of time the Boxer and Thaler results are much more accurate than the Tustin results, and slightly more accurate than the Madwed-Truxal results.

Quantitatively, the mean absolute error of $y(k)$ up to time = 10.0 sec. is approximately .007 for the Tustin results,

Table IV

Machine results for $\ddot{y} + \dot{y} + y = 1$, $\ddot{y}(0)=\dot{y}(0)=y(0)=0$

for three families of Z-Forms, $T=.5$

Time	Exact Solution	Boxer & Thaler	Madwed- Truxal	Tustin
.000	.000000	.000000	.008065	.023810
.500	.104405	.098361	.108221	.129252
1.000	.340300	.335394	.334134	.336573
1.500	.610493	.609910	.601489	.591281
2.000	.849426	.853159	.842477	.826809
2.500	1.023360	1.029590	1.020667	1.005601
3.000	1.124355	1.130785	1.125881	1.115215
3.500	1.161650	1.166437	1.166002	1.161125
4.000	1.153123	1.155320	1.158463	1.158855
4.500	1.118446	1.118023	1.123173	1.127192
5.000	1.074591	1.072182	1.077703	1.083364
5.500	1.033618	1.030177	1.034795	1.040353
6.000	1.002289	.998762	1.001762	1.006041
6.500	.982846	.979951	.981175	.983650
7.000	.974359	.972485	.972201	.972903
7.500	.974152	.973369	.972081	.971411
8.000	.979007	.979144	.977411	.975933
8.500	.985996	.986750	.985050	.983317
9.000	.992934	.993970	.992626	.991065
9.500	.998504	.999536	.998696	.997564
10.000	1.002170	1.003003	1.002669	1.002051

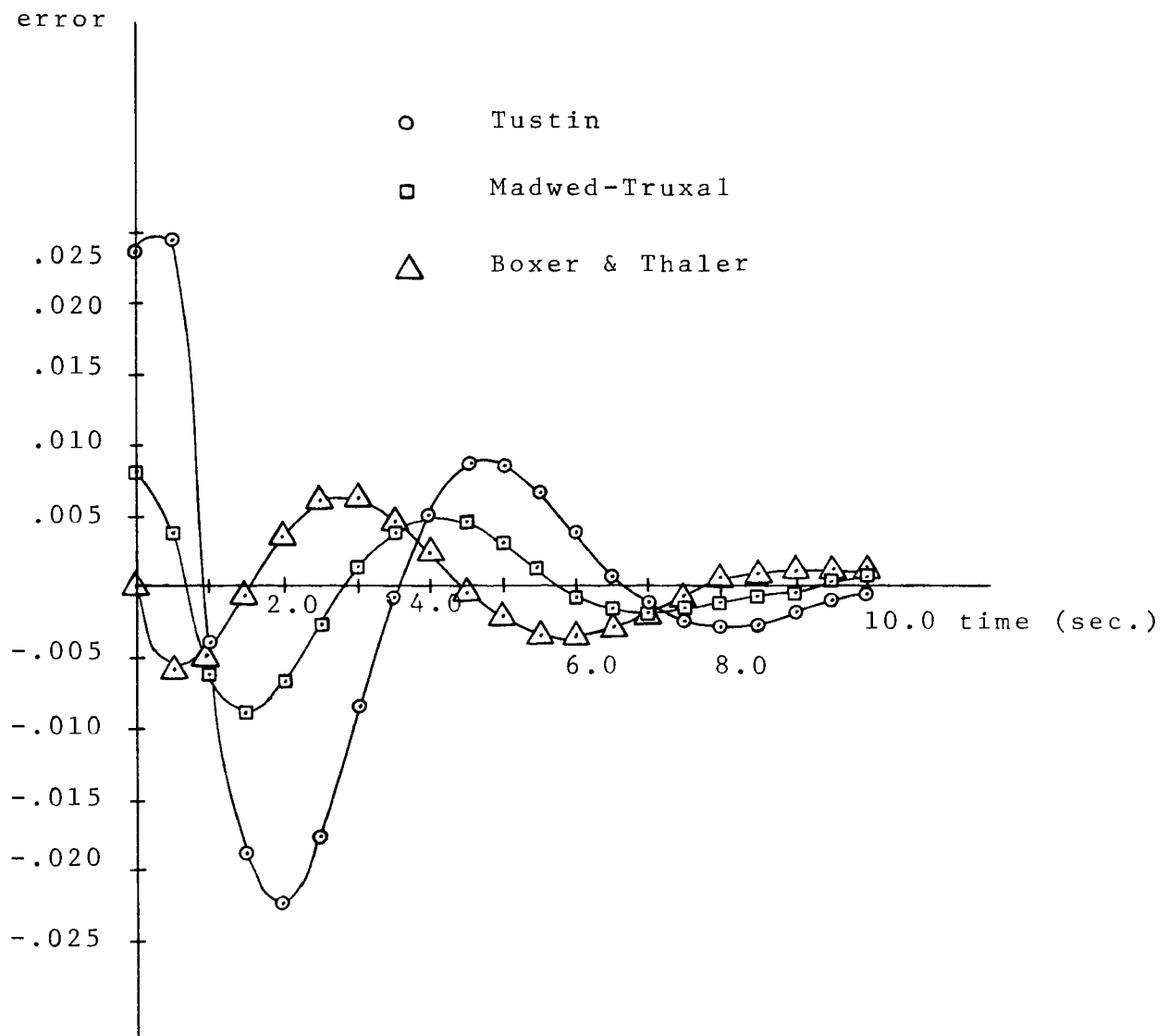


Fig. 1 Error in the solution of
 $\ddot{y} + \dot{y} + y = 1$ for three families of Z-Forms.

and .003 for both the Madwed-Truxal results and the Boxer and Thaler results, using $T=0.5$.

It should be emphasized that, like most practical problems, the preceding example required the use of Z-Forms of several orders. It is always possible that the errors due to the Z-Forms of different orders may tend to cancel at particular points in the discrete-time solution. This cancelling occurs to various extents and at various times depending on the particular problem and the family of Z-Forms used. This is the reason that the Boxer and Thaler results are not the most accurate at some points in the previous example. Furthermore, most of the error in this problem is from the first-order Z-Form, which is the same for all three families.

Boxer and Thaler applied the three families of Z-Forms to the differential equation for simple harmonic motion, namely $\ddot{y} + y = 0$, $y(0)=0$ $\dot{y}(0)=1$. This equation requires only the second-order Z-Form, so that no error cancellation can occur. Their results show a much greater variation in accuracy for the three families of Z-Forms than was observed in the problem of Table I [4].

B. Iteration

The effect of iterating a Z-Form solution to a non-linear differential equation has not been explored in the literature. A non-linear equation is often changed into

a linear one by linearizing, making the problem amenable to solution by applying the Laplace transform [5], [6], [7]. But repeated linearization about successive points, iterating at each point with a method based on the Z-Transform has been barely touched upon in the literature. The iteration technique was merely suggested by Boxer and Thaler [2] and applied to one problem [4]. Hirai suggested a method of trial rather than iteration [8].

No published programs were found for iterating the Z-Form solution of non-linear differential equations. The author wrote his own program, which is listed in the Appendix.

C. New Methods for System Simulation

Fowler [9] and Sage and Smith [10] developed digital simulation techniques which, for a given sampling period T , are more accurate than the Z-Form method of Boxer and Thaler. However, the identification of the discretized system for these newer methods is fairly involved, especially for non-linear problems, compared to mere substitutions for the Z-Form method. In some of these newer methods the discretized system is optimized to give minimum error for a particular type of input, e.g., a step input or a ramp input.

The newer methods are oriented towards systems analysis where the system is given in block diagram form, with a transfer function associated with each block. The system is analyzed without obtaining an over-all transfer function or differential equation, as is required before the Z-Form method can be applied.

D. Programming Z-Transform Inversion

Freeman [1] outlined an algorithm for performing the long division of polynomials to obtain the inverse Z-transform. Crosby and Petersen [11] wrote a complete Fortran program for the long division. Bach [12] suggested a program modification to decrease memory requirements. Boxer and Thaler developed a "modified Z-Form" based on Jury's modified Z-transform for use in obtaining solutions at times between sampling instants [13].

III. Discussion

A. An Alternative Derivation of the Z-Forms Taking into Account Non-Zero Initial Conditions

In a later paper (1957) Boxer [14] gave an alternative derivation of the Boxer and Thaler Z-Forms. The same basic Z-Forms result, but the alternative derivation also indicates how to handle problems with non-zero initial conditions more accurately.

This derivation begins with only one assumption: that an n-th order differential equation can be approximated by an n-th order difference equation. The coefficients are determined which make the difference equation the best approximation.

Assume that the equation $i_1(\tau) = \int y(\tau) d\tau$ is satisfactorily approximated by $a_1 i_1(\tau+T) + a_0 i_1(\tau) = b_1 y(\tau+T) + b_0 y(\tau)$. From Laplace transform theory $y(\tau+nT) = e^{pnT} y(\tau)$ where p is the differentiating operator and variable of the Laplace transform. From the given equation

$$i_1(\tau) = \frac{1}{p} y(\tau)$$

Therefore

$$a_1 e^{pT} i_1(\tau) + a_0 i_1(\tau) = b_1 e^{pT} y(\tau) + b_0 y(\tau)$$

$$a_1 e^{pT} \left(\frac{1}{p}\right) y(\tau) + a_0 \left(\frac{1}{p}\right) y(\tau) = b_1 e^{pT} y(\tau) + b_0 y(\tau)$$

Expand e^{pT} in a Taylor series about the origin.

$$\begin{aligned}
& a_1 \left(1 + pT + \frac{(pT)^2}{2} + \frac{(pT)^3}{6} + \dots \right) \left(\frac{1}{p} \right) y(\tau) + a_0 \left(\frac{1}{p} \right) y(\tau) \\
& = b_1 \left(1 + pT + \frac{(pT)^2}{2} + \frac{(pT)^3}{6} + \dots \right) y(\tau) + b_0 y(\tau)
\end{aligned}$$

Equating the coefficients of like powers of p yields the following three equations.

$$a_1 + a_0 = 0$$

$$a_1 T = b_1 + b_0$$

$$a_1 \frac{T^2}{2} = b_1 T$$

One of the four variables may be assumed to have any arbitrary value. If we choose $a_0=1$ and solve the three equations simultaneously we obtain

$$a_1 = -1$$

$$b_0 = \frac{T}{2}$$

$$b_1 = \frac{T}{2}$$

Substitution into the difference equations yields

$$i_1(\tau+T) - i_1(\tau) = \frac{T}{2} [y(\tau+T) + y(\tau)]$$

Note that this equation is the trapezoidal rule of integration. Taking the Z-Transform of this equation gives us the first-order Z-Form.

$$zI_1(z) - zi_1(0) - I_1(z) = \frac{T}{2} [zY(z) - zy(0) + Y(z)]$$

$$I_1(z) = \frac{T(z+1)}{2(z-1)} Y(z) - \frac{Tz}{z-1} y(0) + \frac{z}{z-1} i_1(0) \quad (4)$$

If initial conditions are zero (4) reduces to

$$I_1(z) = \frac{T(z+1)}{2(z-1)} Y(z)$$

$$\frac{I_1(z)}{Y(z)} = \frac{T(z+1)}{2(z-1)}$$

$$\frac{s^{-1}Y(z)}{Y(z)} = \frac{T(z+1)}{2(z-1)}$$

$$s^{-1} = \frac{T(z+1)}{2(z-1)}$$

which is the first-order Z-Form derived previously.

The second term on the right-hand side of (4) involves the initial condition of $y(t)$. The last term on the right-hand side of (4) involves the initial value of the integral of $y(t)$, and may normally be neglected in practical problems.

The modified second-order Z-Form can be derived in a similar manner. Beginning with the equation

$$i_2(\tau) = \iint y(\tau) d\tau d\tau$$

the difference equation approximation obtained is

$$i_2(\tau+2T) - 2i_2(\tau+T) + i_2(\tau) = \frac{T^2}{12} [y(\tau+2T) + 10y(\tau+T) + y(\tau)]$$

Taking the Z-Transform yields the second-order Z-Form.

$$\begin{aligned} z^2 [I_2(z) - i_2(0) - z^{-1}i_2(T)] - 2z [I_2(z) - i_2(0)] + I_2(z) \\ = \frac{T^2}{12} \{ z^2 [Y(z) - y(0) - z^{-1}y(T)] + 10z [Y(z) - y(0)] + Y(z) \} \end{aligned}$$

$$\begin{aligned} I_2(z) [z^2 - 2z + 1] - z^2 i_2(0) - z i_2(T) + 2z i_2(0) \\ = \frac{T^2}{12} Y[z^2 + 10z + 1] - \frac{T^2 z^2}{12} y(0) - \frac{T^2 z}{12} y(T) - 10z y(0) \end{aligned}$$

The terms containing $i_2(T)$ and $y(T)$ can be more usefully expressed in terms of $i_2(0)$ and $y(0)$ by use of Taylor series expansions

$$y(T) = y(0) + T\dot{y}(0) + \frac{T^2}{2!}\ddot{y}(0) + \dots$$

$$i_2(T) = i_2(0) + Ti_2'(0) + \frac{T^2}{2!}i_2''(0) + \dots$$

Noting that $i_2(0) = y^{(-2)}(0)$,

$$i_2(T) = y^{(-2)}(0) + Ty^{(-1)}(0) + \frac{T^2}{2!}y(0) + \dots$$

After substitution and algebraic manipulation on the second-order Z-Form is obtained in the following form:

$$\begin{aligned} I_2(z) = & \frac{T^2(z^2+10z+1)}{12(z-1)^2}Y(z) - \frac{T^2z(z+5)}{12(z-1)^2}y(0) \\ & + \frac{T^3z}{12(z-1)^2}\dot{y}(0) - \frac{T^5z}{180(z-1)^2}\ddot{y}(0) + \dots \\ & + \frac{Tz}{(z-1)^2}y^{(-1)}(0) + \frac{z}{z-1}y^{(-2)}(0) \end{aligned} \quad (5)$$

where it should be noted that the second derivative term is absent. Higher-order Z-Forms can be obtained in a similar manner.

To illustrate the use of the Z-Forms, including initial condition terms, an example problem will be worked out in detail.

$$\ddot{y} + y = 0, \quad y(0)=0 \quad \dot{y}(0)=1$$

Taking the Laplace transform yields

$$s^2 Y(s) - sy(0) - \dot{y}(0) + Y(s) = 0$$

$$Y(s) + \frac{1}{s^2} Y(s) = \frac{1}{s^2}$$

Taking the Z-Transform of this equation by using the first three terms of (5) for the integrating operator $\frac{1}{s^2}$

yields

$$Y(z) + \frac{T^2(z^2 + 10z + 1)}{12(z-1)^2} Y(z) - \frac{T^2 z(z+5)}{12(z-1)^2} Y(0) + \frac{T^3 z}{12(z-1)^2} \dot{Y}(0) = \frac{Tz}{(z-1)^2}$$

The terms in (5) containing higher-order initial conditions cannot be used in this problem because the higher-order initial conditions are not given. The neglect of these terms will have negligible effect on the solution if T is chosen small.

Substituting $y(0)=0$ and $\dot{y}(0)=1$ yields

$$Y(z) + \frac{T^2(z^2 + 10z + 1)}{12(z-1)^2} Y(z) + \frac{T^3 z}{12(z-1)^2} = \frac{Tz}{(z-1)^2}$$

After some algebraic manipulation we obtain

$$Y(z) = \frac{[T - \frac{T^3}{12}]z}{[1 + \frac{T^2}{12}]z^2 + [-2 + \frac{10T^2}{12}]z + [1 + \frac{T^2}{12}]}$$

Machine results for this solution are presented in Table V, along with the exact solution, $y(t)=\sin(t)$. The sampling period T was chosen to be 0.2. Note that there is no phase shift in the Z-Form solution.

This method of substituting the Z-Forms without first solving explicitly for $Y(s)$ can also be applied to problems with zero initial conditions, of course. If this modified method is applied to the equation $\dot{y}+2y=1$ $y(0)=0$ the results obtained are slightly different from those obtained previously in this paper. Table VI contains the results obtained by the two methods. Note the superior results obtained by the modified method for $t=0.0$, 0.1 , and 0.2 . For later times the results are of comparable accuracy.

As demonstrated in the preceding problem, the modified Z-Form method generally leads to superior accuracy for the first few instants of discrete-time. In the modified method exact Z-transforms are used for constants and functions of the independent variable in the equation, rather than Z-Form approximations. The Z-Forms are used only for those terms containing the dependent variable. This fact probably accounts for the superiority of the modified method for the first few terms.

Table V

Machine results for $\ddot{y}+y=0$, $y(0)=0$ $\dot{y}(0)=1$.

Time	Exact Solution	Z-Form Solution, T=.2
.000	.000000	.000000
.200	.198669	.198671
.400	.389418	.389422
.600	.564642	.564647
.800	.717356	.717362
1.000	.841471	.841478
1.200	.932039	.932046
1.400	.985450	.985456
1.600	.999574	.999579
1.800	.973848	.973852
2.000	.909297	.909300
2.200	.808496	.808497
2.400	.675463	.675461
2.600	.515501	.515497
2.800	.334988	.334981
3.000	.141120	.141111
3.200	-.058374	-.058385
3.400	-.255541	-.255553
3.600	-.442520	-.442534
3.800	-.611858	-.611871
4.000	-.756802	-.756815
4.200	-.871576	-.871587
4.400	-.951602	-.951612
4.600	-.993691	-.993698
4.800	-.996165	-.996169
5.000	-.958924	-.958925
5.200	-.883455	-.883452
5.400	-.772764	-.772757
5.600	-.631267	-.631256
5.800	-.464602	-.464588
6.000	-.279416	-.279398
6.200	-.083089	-.083069
6.400	.116549	.116571
6.600	.311541	.311564
6.800	.494113	.494136
7.000	.656987	.657008
7.200	.793668	.793687
7.400	.898708	.898724
7.600	.967920	.967931
7.800	.998543	.998550
8.000	.989358	.989360
8.200	.940731	.940727
8.400	.854599	.854589

continued

Table V (Continued)

Time	Exact Solution	Z-Form Solution, T=.2
8.600	.734397	.734382
8.800	.584917	.584897
9.000	.412118	.412094
9.200	.222890	.222861
9.400	.024775	.024744
9.600	-.174327	-.174359
9.800	-.366479	-.366511
10.000	-.544021	-.544052

Table VI

Machine results for $\dot{y}+2y=1$, $y(0)=0$, by both original and modified methods.

Time	Exact Solution	Original Z-Form Solution T=.1	Modified Solution T=.1
.000	.000000	.000576	.000000
.100	.090635	.089532	.090909
.200	.164840	.164162	.165289
.300	.225594	.225224	.226146
.400	.275336	.275183	.275937
.500	.316060	.316059	.316676
.600	.349403	.349503	.350008
.700	.376702	.376866	.377279
.800	.399052	.399254	.399592
.900	.417351	.417571	.417848
1.000	.432332	.432558	.432785
1.100	.444598	.444820	.445006
1.200	.454641	.454853	.455005
1.300	.462863	.463062	.463186
1.400	.469595	.469778	.469879
1.500	.475106	.475273	.475356
1.600	.479619	.479769	.479836
1.700	.483313	.483447	.483503
1.800	.486338	.486457	.486502
1.900	.488815	.488919	.488956
2.000	.490842	.490934	.490964
2.100	.492502	.492582	.492607
2.200	.493861	.493931	.493951
2.300	.494974	.495034	.495051
2.400	.495885	.495937	.495951
2.500	.496631	.496676	.496687
2.600	.497242	.497280	.497289
2.700	.497742	.497775	.497782
2.800	.498151	.498179	.498185
2.900	.498486	.498510	.498515
3.000	.498761	.498781	.498785
3.100	.498985	.499003	.499006
3.200	.499169	.499184	.499187
3.300	.499320	.499332	.499335
3.400	.499443	.499454	.499456
3.500	.499544	.499553	.499555
3.600	.499627	.499634	.499636
3.700	.499694	.499701	.499702
3.800	.499750	.499755	.499756
3.900	.499795	.499800	.499800
4.000	.499832	.499836	.499837

B. Error

Two sources of error in Z-Form solutions will be discussed. In addition there are round-off and significance errors inherent in machine computations. It will be assumed that enough significant digits are carried in the computations to assure that these errors are negligible compared to other errors.

1. Discretization Error

Discretization error is due to the truncation of the Laurent series used in deriving the Z-Forms. From the viewpoint of the difference equation derivation this error arises because an n -th order difference equation is an imperfect approximation for an n -th order differential equation. This type of error approaches zero as the sampling period T approaches zero.

Jury said that there exists no general formula for the upper bound of the error [13]. To attain a given accuracy in a particular problem Jury and others suggested solving the problem repeatedly and successively subdividing T . When the last two solutions are identical to the specified number of decimal places, the solution is assumed to meet accuracy requirements. The present author found both theoretical and experimental justification for some general quantitative statements about Z-Form solution error.

Boxer and Thaler expanded each of their Z-Forms in an infinite series in s about $s=0$ [14]. Their results are shown in Table VII. The terms beyond the first terms represent error. Boxer and Thaler showed how these error terms for the different ordered Z-Forms combine in a unique way for each problem to produce error in the Z-Form solution of a differential equation. They achieved close agreement between theoretical error and calculated error.

From Table VII it can be seen that, for small T , the error of the first-order Z-Form is proportional to T^2 ; the error of the second and third-order Z-Forms proportional to T^4 , etc. These error relationships were verified in several computer runs.

Consider the equation $\ddot{y} + y = 0$, $y(0)=0$, $\dot{y}(0)=1$ for example. In this problem only the Z-Form for s^{-2} is used. If the Z-Form for s^{-1} had been required, an error of about 0.3% per second would be expected for about 30 samples per cycle, as calculated in the Introduction. But examination of the results in Table V reveals an error accumulation rate of about 0.0006% per second, averaged over the first 10.0 seconds. This figure agrees closely with the theoretical accuracy of the second-order Z-Form.

For the first-order Z-Form the discretization error is proportional to T^2 . This is the same error-to-step-size relationship as in trapezoidal rule integration. This

Table VII

Expansion of the Boxer and Thaler Z-Forms about $s=0$.

$\frac{1}{s^n}$	Z-Form	Series Expansion
$\frac{1}{s}$	$\frac{T(z+1)}{2(z-1)}$	$\frac{1}{s} + \frac{sT^2}{12} - \frac{s^3T^4}{720} + \dots$
$\frac{1}{s^2}$	$\frac{T^2(z^2+10z+1)}{12(z-1)^2}$	$\frac{1}{s^2} + \frac{s^2T^4}{240} - \frac{s^4T^6}{6048} + \dots$
$\frac{1}{s^3}$	$\frac{T^3z(z+1)}{2(z-1)^3}$	$\frac{1}{s^3} - \frac{sT^4}{240} + \frac{s^3T^6}{3024} + \dots$
$\frac{1}{s^4}$	$\frac{T^4z(z^2+4z+1)}{6(z-1)^4} - \frac{T^4}{720}$	$\frac{1}{s^4} - \frac{s^2T^6}{3024} + \dots$

similarity is not surprising since the first-order Z-Form is equivalent to trapezoidal rule integration.

Consider the equation $\dot{y}=e^t$, $y(0)=1$.

The solution of course is $y(t)=e^t$.

The results of Z-Form solutions for various values of T are contained in Table VIII, together with the exact solution. The error using Z-Forms is just what one would expect using the trapezoidal rule to integrate $y=e^t$, $y(0)=1$ with the same value for T . Specifically, the error of trapezoidal rule integration of $f(t)$ is given by [15]

$$e = \frac{-(b-a)}{12} T^2 f''(\xi) \quad a < \xi < b$$

where e = the difference between the true solution and the approximation

b = the upper limit of integration

a = the lower limit of integration

T = the interval size or sampling period

2. Recursion Error

Another type of error, recursion error, occurs only in the case of non-linear differential equations. It is caused by the use of the last calculated value of a solution to compute the next value of the solution, when in fact the value of the solution one is seeking should be used, an obvious impossibility.

The magnitude of recursion error varies greatly from problem to problem, depending upon the form of the coefficients in the ratio of polynomials obtained after

Table VIII

Machine results for $\dot{y}=e^t$, $y(0)=1$.

Z-Form Solutions

Time	Exact Solution	T=.5	T=.25	T=.1	T=.05
.000	1.000000	1.000000	1.000000	1.000000	1.000000
.500	1.648721	1.666667	1.653061	1.649409	1.648893
1.000	2.718282	2.777778	2.732611	2.720551	2.718848
1.500	4.481689	4.629630	4.517174	4.487303	4.483090
2.000	7.389056	7.716049	7.467165	7.401400	7.392137
2.500	12.182494	12.860082	12.343681	12.207939	12.188843
3.000	20.085537	21.433470	20.404860	20.135890	20.098099
3.500	33.115452	35.722451	33.730484	33.212326	33.139617
4.000	54.598150	59.537418	55.758555	54.780724	54.643686
4.500	90.017131	99.229029	92.172304	90.355843	90.101596
5.000	148.413160	165.381720	152.366460	149.033780	148.567900

substituting the Z-Forms. In a given problem the magnitude of the recursion error which occurs in an interval T is directly related to the magnitude of the first derivative of the dependent variable at some point in the interval, and also to the sampling period T . This relationship is evident upon detailed examination of the long division process.

The first term of the denominator polynomial of $y(z)$ is normally of the form $A+BC^n$, where A and B are constant (for a particular T), C represents the last calculated value of the discrete-time solution $Y(k)$, and n is the degree of the equation minus one. This is the term which is successively divided into the remainder polynomials to obtain the discrete-time solution. If the magnitude of the product BC^n is significant relative to the magnitude of A , the recursion error may still be small if C does not change much over the interval T . To say it another way, the recursion error may still be small if the magnitude of the first derivative of $y(t)$ is small.

The approximate solution diverges from the true solution when the term $A+BC^n$ is very small compared to its dividend. This situation can sometimes be avoided or postponed by a wise choice of T , because A and/or B are functions of T .

Huggins [16] showed that error in the first term of the denominator can cause the quotient sequence to diverge from the true solution when the first term of the

denominator is much smaller than the other denominator terms. In the non-linear problems discussed here this type of error arises from the use of an approximate value for C. In many practical problems with empirically determined coefficients there may be additional error in the first denominator term of $Y(z)$. Huggins also proposed a type of "smoothing" operation to minimize the error due to an inaccurate first term in the denominator.

3. Convergence of Iteration

It was demonstrated in the Introduction that iteration can improve the accuracy of the Z-Form solution of a non-linear differential equation. This improvement can occur only if the iteration converges. A condition sufficient to insure convergence will now be presented.

The discrete-time solution at a point is obtained by dividing the first term of the denominator polynomial of $Y(z)$ into the first term of the remainder. Mathematically this may be expressed

$$\frac{R}{A+BCn} = y(k)$$

where R is the coefficient of the first term in the remainder. The iteration process is equivalent to the solution by linear iteration of

$$\frac{R}{A+B[y(k)]^n} = y(k) \quad (6)$$

where the initial approximation of $y(k)$ is taken as $y(k-1)$ or C.

This iteration may be expressed by a general recursion formula

$$x_{i+1} = g(x_i) \quad i=0,1,\dots$$

A theorem from numerical analysis [15] states that, to assure convergence, the inequality $|g'(\xi)| < 1$, where ξ is the true value of the solution, must hold. In addition the initial guess to x_i must be in an interval where $|g'(x)| \leq K < 1$, and $g(x)$ and $g'(x)$ are continuous.

To apply this theorem to (6), the derivative of the left-hand side with respect to $h(k)$ must be found. Iteration converges if and only if the absolute value of this derivative is less than 1. This convergence condition is expressed mathematically as

$$\left| \frac{-RnB[y(k)]^{n-1}}{(A+B[y(k)]^n)^2} \right| < 1 \quad (7)$$

An example problem will be worked out for which iteration does not converge. Consider the non-linear differential equation

$$y^2 \ddot{y} + y = 0, \quad y(0) = 1 \quad \dot{y}(0) = 0$$

Taking the Laplace transform leads to

$$C^2 Y(s) + \frac{1}{s^2} Y(s) = \frac{C^2}{s}$$

Taking the Z-Transform using the Boxer and Thaler Z-Forms for the second-order integrating operator yields, after algebraic manipulation

$$Y(z) = \frac{[C^2 + \frac{T^2}{12}]z^2 + [-C^2 + \frac{5T^2}{12}]z}{[C^2 + \frac{T^2}{12}]z^2 + [-2C^2 + \frac{10T^2}{12}]z + [C^2 + \frac{T^2}{12}]}$$

Since the initial value of $y(t)$ is one, it can be assumed that the value of the solution at the next point in discrete-time is near one, provided T is chosen small enough and there are not discontinuities in the first interval. If T is chosen equal to 0.2, the left-hand side of the convergence condition of (7) becomes approximately

$$\left| \frac{(-1)(2)(1)(1)^{2-1}}{(.003 + 1^2)^2} \right| \text{ where } y(0) \text{ has been substituted for}$$

$y(k)$ which is clearly greater than 1. Therefore the iteration does not converge. Machine results for this problem, using various numbers of iterations, are given in Table IX.

Note that the solution appears not very good even without iteration, becoming completely worthless beyond time equal to 1.4. But iteration makes matters worse for all values of time.

Results from the preceding problem and other problems suggest the general conclusion that if iteration does not converge, then the results obtained without iterating are likely to be poor, if usable at all.

Table IX

Machine results for $y'' + y = 0$, $y(0) = 1$ $\dot{y}(0) = 0$ Z-Form Solutions, $T = .2$

Time	No Iteration	One Iteration	Multiple Iterations
.000	1.000000	1.000000	1.000000
.200	.980066	1.061149	10.845514
.400	.958777	1.270297	660.225010
.600	.895933	2.795071	-7147.935700
.800	.860000	60.157000	70819.132000
1.000	.707189	-555.383320	
1.200	.654119	5331.606300	
1.400	.248310	-52760.624000	
1.600	-1.924919	522274.630000	
1.800	-.075575	-5169985.600000	
2.000	-47.809815	51177581.000000	

C. Other Difficulties and Limitations of Iteration

Iteration is most effective when applied to second-degree differential equations non-linear in y , rather than \dot{y} or \ddot{y} . The author had some success in applying the Z-Form method to equations non-linear in \dot{y} , such as a particular form of the Ricatti equation

$$\ddot{y} + 0.1 \dot{y}^2 = 32.2 \quad y(0)=\dot{y}(0)=0$$

which describes the motion of a body falling from rest through a viscous fluid. The results of the Z-Form solution are presented in Table X. A second-order Newton backward-difference formula was used to calculate $\dot{y}(k)$.

Iteration of the Z-Form solution of this equation, and others non-linear in \dot{y} , had a more complicated convergence condition than that given for equations non-linear in y . A major difficulty is the accurate calculation of \dot{y} . Since numerical differentiation is basically an unstable process, the iteration of equations non-linear in \dot{y} is very likely to be divergent. For first-order equations non-linear in \dot{y} it is possible, and probably advisable, to calculate \dot{y} by direct functional evaluation.

Higher-degree non-linear equations require squaring or cubing of the discrete-time solution at successive points. This process magnifies any error which is present, and propagates it to the solution at the next point.

The Z-Form method was applied to Van der Pol's equation

Table X

Machine results for $\ddot{y} + 0.1\dot{y}^2 = 32.2$, $y(0) = \dot{y}(0) = 0$

Z-Form Solutions, $T = .1$

Time	One Iteration	Multiple Iterations
.000	.000000	.000000
.100	.161000	.161000
.200	.628972	.629267
.300	1.367161	1.368908
.400	2.327373	2.332847
.500	3.442868	3.460224
.600	4.679687	4.712326
.700	6.018411	6.068376
.800	7.431468	7.503053
.900	8.901927	8.999282
1.000	10.413724	10.542456
1.100	11.954303	12.120324
1.200	13.510385	13.718008
1.300	15.065638	15.299891
1.400	16.594660	16.662327
1.500	18.048451	15.876214
1.600	19.323151	9.228775
1.700	20.181358	6.784383
1.800	20.106204	9.734340
1.900	18.215107	12.204497
2.000	14.048366	14.643519

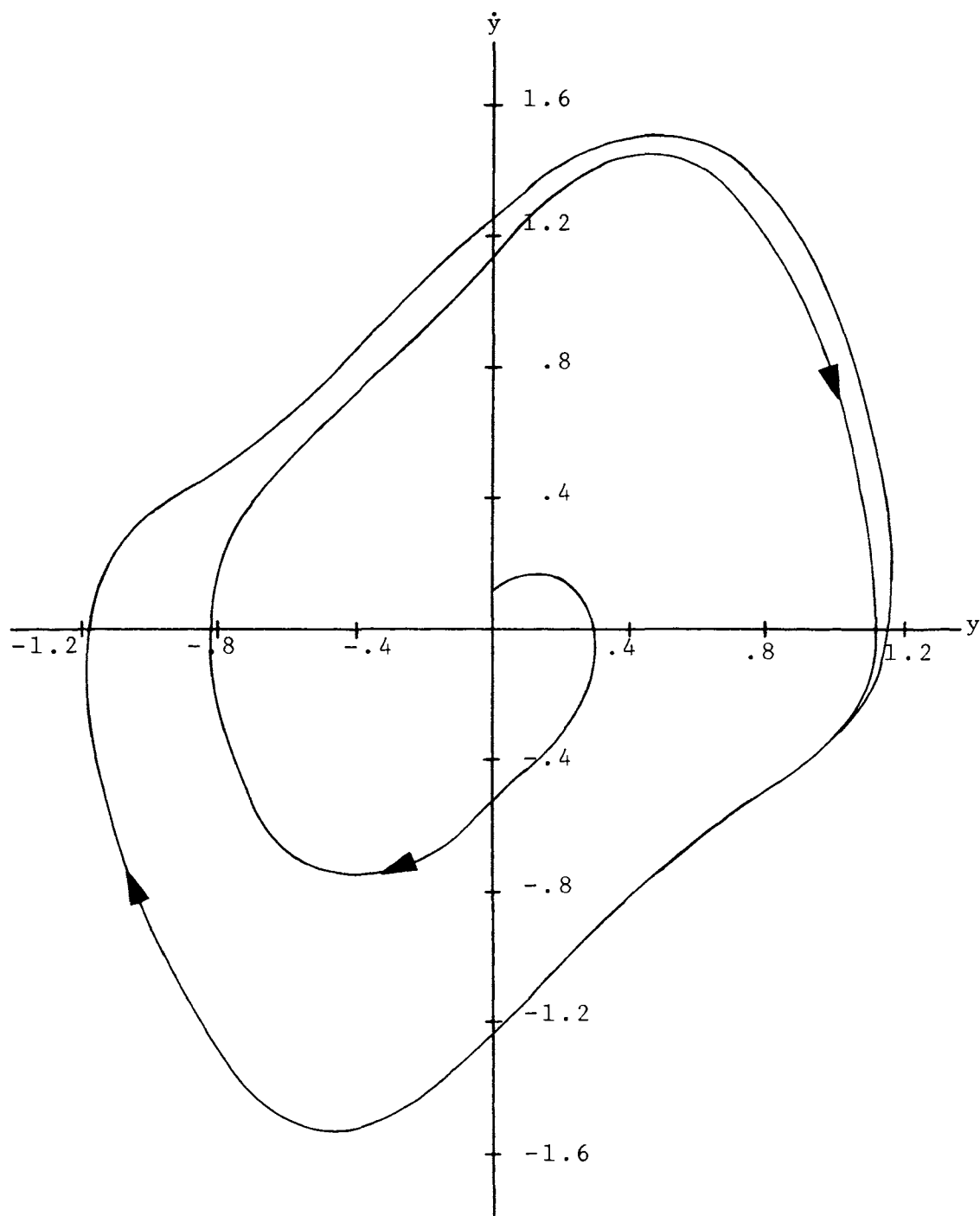


Fig. 2 Solution of Van der Pol's Equation.

of the form

$$\ddot{y} + (1 - y^2)\dot{y} + y = 0, \quad y(0) = 0 \quad \dot{y}(0) = 0.1$$

with fair results. After the initial perturbation represented by $\dot{y}(0) = 0.1$, the solution fell into a limit cycle, as expected. The general shape of this limit cycle (see Figure 2) is the same as that in published texts [17], [18], [19] but the magnitude is inexplicably smaller.

D. Choice of Sampling Period

Most differential equations require the use of the first-order Z-Form, which is the least accurate of the Z-Forms. Therefore the error in a Z-Form solution may be assumed to be dominated by the error due to the first-order Z-Form.

The problem of choosing a suitable sampling period T is clarified by considering the close relationship between the Z-Form for s^{-1} and trapezoidal rule integration. Salzer [20] plotted the accuracy of trapezoidal rule integration versus the ratio of sampling period to the period of the function to be integrated. This plot is reproduced in Figure 3. Note that trapezoidal rule integration attenuates at higher frequencies, thereby counteracting the effect of high-frequency noise.

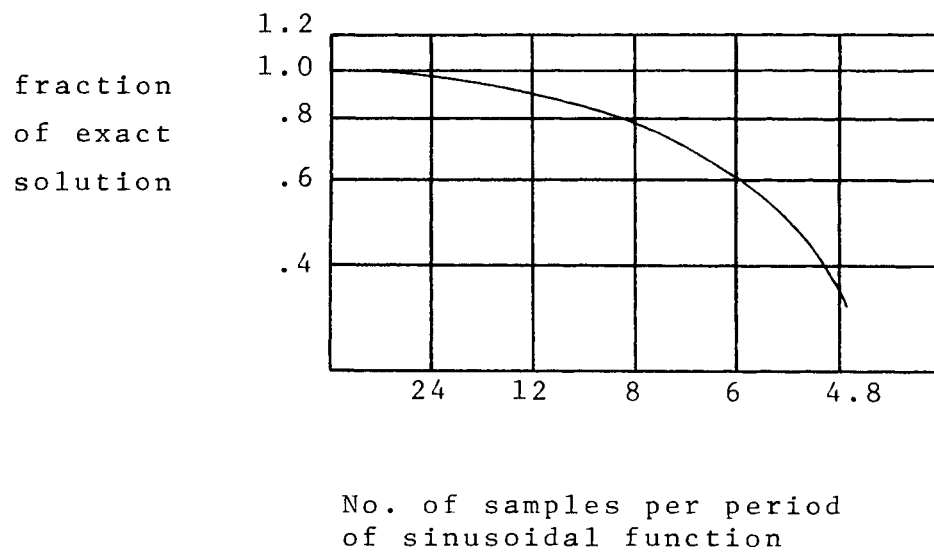


Fig. 3 Accuracy of trapezoidal rule integration

E. Notes on the Computer Program

The core of the program is based on Freeman's algorithm for the division of two polynomials, specifically the ratio of two polynomials representing the Z-Transform of a discrete-time function. The program was written to be useful in a wide variety of problems. User-specified parameters are listed below.

- 1) Type of equation
 - a. Linear, time-stationary
 - b. Linear, time-variant
 - c. Non-linear in y
 - d. Non-linear in \dot{y}
- 2) Number of zero terms before the first non-zero term in $y(k)$
- 3) Iterated or non-iterated solution (for non-linear equations only)
 - a. Max. no. of iterations
 - b. Relative error convergence criterion
- 4) Endpoint (max. value of independent variable)
- 5) Max. no. of solution points to be printed
- 6) Sampling period T

No attempt was made to economize on memory requirements or computation time. A maximum of 201 points of discrete-time solution may be calculated and printed. The corresponding first derivative of the solution at each point is also printed, to aid in making a phase-plane plot.

The statements which specify the form of the numerator and denominator coefficients must be changed for each problem. Other characteristics of the program are described in the comment statements of the program listing in the Appendix.

IV. Conclusions

The Z-Form method is an easy and systematic procedure for obtaining numerical solutions to some ordinary differential equations, both linear and non-linear. One should be cautious in applying the method, however, because it breaks down completely in some problems. If the method works at all in solving a non-linear differential equation, it is usually worth the extra effort to iterate each stage of the long division once to obtain greater accuracy. This is especially true if it can be established in advance that the iteration converges throughout the region of interest.

Iteration often reduces recursion error; a small sample period T reduces discretization error. These errors may be additive or cancelling.

Though discretization error and recursion error were discussed separately, they are not independent. This means that a reduction in T may reduce error, and iteration with the original T may also reduce error, but if both a smaller T and iteration are employed the total error reduction is less than the sum of the individual error reductions. Iteration is often the more efficient way of reducing error.

In some problems with periodic solutions it seems likely that the net recursion error over one period would

Table XI

Numbers of arithmetic operations required by different methods in various problems.

	Fourth-order Runge-Kutta	Second-order Runge-Kutta	Z-Form (No Iteration)
$\dot{y} + y^2 = 1$			
Multiplications & divisions	13	4	5
<u>Additions & subtractions</u>	15	4	4
$\dot{y} = y + t^2$			
Multiplications & divisions	13	4	5
<u>Additions & subtractions</u>	15	4	4
$\ddot{y} + \dot{y} + y = 1$			
Multiplications & divisions	33	6	4
<u>Additions & subtractions</u>	32	10	3
$\ddot{y} + y^2 = 0$			
Multiplications & divisions	22	8	6
<u>Additions & subtractions</u>	14	6	5

Note: Numbers indicate operations required per step in discrete-time.

be close to zero. The recursion error would not build up from one period to the next, even though the solution at any particular point might have substantial error. The recursion errors at the different points are likely to appear as a phase shift in the periodic solution.

Iteration could probably improve solution accuracy in both magnitude and phase, but this is an hypothesis needing further verification.

Table XI compares the Z-Form method and two Runge-Kutta methods in numbers of arithmetic operations required to solve several differential equations. Machine time requirements are directly related to number of arithmetic operations. If iteration is used with the Z-Form method, one extra division and one denominator coefficient evaluation must be performed for each iteration.

For solving differential equations the Z-Form method cannot compete with Runge-Kutta or predictor-corrector methods in either accuracy or reliability. Machine time requirements may be substantially less for the Z-Form method however.

The Z-Form method is a systematic way of converting a continuous transfer function into an approximately equivalent discrete-time transfer function. This feature could make the Z-Forms useful in digital filter synthesis and in the analysis and design of sampled-data systems.

V. Appendix - Listing of the Computer Program

```

100C THIS IS A GENERAL PROGRAM TO FIND THE INVERSE Z-TRANSFORM
105C OF A Y(Z) GIVEN AS THE RATIO OF 2 POLYNOMIALS IN POWERS
110C OF Z**-1. FIRST THE NUMBERS OF TERMS IN THE NUMERATOR
115C AND DENOMINATOR ARE INPUT. THEN THE NUMBER OF DELAY
120C PERIODS AFTER TIME=0 BEFORE THE FIRST NON-ZERO Y(K) IS
125C INPUT. THEN AN INDICATION OF THE MEANING OF "C" IN THE
130C A'S AND B'S IS INPUT. IF C REPRESENTS Y OR Y', THEN THE
135C INITIAL VALUE OF C, MAXIMUM NUMBER OF ITERATIONS, AND THE
140C RELATIVE ERROR CRITERION ARE INPUT. FINALLY THE END-
145C POINT, MAXIMUM NUMBER OF TERMS TO BE PRINTED, AND SAMPLE
150C PERIOD ARE INPUT.
155C THE PROGRAM PERFORMS THE LONG DIVISION AND PRINTS THE
160C COEFFICIENTS OF THE QUOTIENT TERMS AS THE DISCRETE TIME
165C SOLUTION Y(K). THE FIRST DERIVATIVE OF Y(K) IS ALSO
170C CALCULATED AND PRINTED, TO FACILITATE A PHASE PLANE PLOT.
175     DIMENSION A(10),B(10),BD(10,10),Y(201),YPRM(201),TIME
           (201)
176     DIMENSION IT(201)
180     COMMON A,B,T
200     PRINT,"NO. OF TERMS IN NUMERATOR"
205     INPUT,NA
210     PRINT,"NO. OF TERMS IN DENOMINATOR"
215     INPUT,NB
220     NAA=NA+1
225     NBB=NB+1
230     DO 235 I=NAA,10
235 235 A(I)=0.
240     DO 245 I=NBB,10
245 245 B(I)=0.
250     PRINT,"NO. OF DELAY PERIODS"
255     INPUT,KDLAY
260     PRINT,"IF C DOES NOT APPEAR IN THE A'S AND B'S.
           TYPE 1."
265     PRINT,"IF C REPRESENTS TIME, TYPE 2."
270     PRINT,"IF C REPRESENTS Y, TYPE 3."
275     PRINT,"IF C REPRESENTS Y', TYPE 4."
280     INPUT,MC
285 285 GO TO (320,320,290,290)MC
290 290 PRINT,"INITIAL VALUE OF C"
295     INPUT,CO
300     PRINT,"MAX. NO. OF ITERATIONS"
305     INPUT,MAXIT
310     PRINT,"RELATIVE ERROR CRITERION"
315     INPUT,E
320 320 PRINT,"ENDPOINT"

```

Appendix (Continued)

```

325      INPUT,END
330      PRINT,"MAX. NO. OF TERMS TO BE PRINTED"
335      INPUT,MAXTM
340 340  PRINT,"T"
345      INPUT,T
346      DO 348 I=1,10
347      DO 348 J=1,10
348 348  BE (I,J)=0.
350      IF (KDLAY) 370,370,355
355 355  DO 365 K=1,KDLAY
360      Y(K)=0.
365 365  YPRM(K)=0.
370 370  TIME(1)=0.
371      M=END/T+1
375      DO 380 K=2,M
380 380  TIME(K)=T+TIME(K-1)
385      DO 390 K=1,M
390 390  IT(K)=0
395      CONTINUE
400      GO TO (420,405,415,415)MC
405 405  C=TIME(KDLAY+1)
410      GO TO 420
415 415  C=CO
420 420  CALL COEF(C)
425      Y(KDLAY+1)=Y(KDLAY+1)/T
430      YPRM(KDLAY+1)=Y(KDLAY+1)/T
435      DO 450 I=1,NB
440      DO 450 J=1,9
445      N=11-J
450 450  BD(I,N)=BD(I,N-1)
455      DO 460 I=1,NB
460 460  BD(I,1)=B(I)
465      GO TO (495,470,480,490)MC
470 470  C=TIME(KDLAY+2)
475      GO TO 495
480 480  C=Y(KDLAY+1)
485      GO TO 495
490 490  C=YPRM(KDLAY+1)
495 495  K3=2+KDLAY
500 500  NBK=NB+KDLAY
510      DO 635 K=K3,NBK
511      KB=K-KDLAY
515      GO TO (525,525,520,520)MC
520 520  DO 600 I=1,MAXIT
525 525  CALL COEF(C)
530      SUM=0.
535      DO 545 J=2,KB
540      L=K-J
545 545  SUM=SUM+BD(J,J-1)*Y(L+1)/B(1)
550      Y(K)=A(KB)/B(1)-SUM

```

Appendix (Continued)

```

555      YPRM(K)=Y(K)-Y(K-1))/T
560      GO TO (610,607,565,566)MC
565 565 IF (ABS(Y(K)-C)-E*ABS(YK))) 570,570,580
566 566 IF (ABS(YPRM(K)-C)-E*ABS(YPRM(K))) 570,570,580
570 570 IT(K)=I
571      GO TO (500,500,572,574)MC
572 572 C=Y(K)
573      GO TO 610
574 574 C=YPRM(K)
575      GO TO 610
580 580 GO TO (500,500,585,595)MC
585 585 C=Y(K)
590      GO TO 600
595 595 C=YPRM(K)
600 600 CONTINUE
605      IT(K)=MAXIT
606      GO TO 610
607 607 C=TIME(K+1)
610 610 DO 625 I=1,NB
615      DO 625 J=1,9
620      N=11-J
625 625 BD(I,N)=BD(I,N-1)
630      DO 635 I=1,NB
635 635 BD(I,1)=B(I)
640      NBK1=NBK+1
645      DO 780 K=NBK1,M
650      GO TO (660,660,655,655)MC
655 655 DO 735 I=1,MAXIT
660 660 CALL COEF(C)
665      SUM=0.
670      DO 680 J=2,NB
675      L=K-J
680 680 SUM=SUM+BD(J,J-1)*Y(L+1)/B(1)
685      Y(K)=-SUM
690      YPRM(K)=(3.*4.8Y(K-1)+Y(K-2))/(2.*T)
695      GO TO (755,750,700,701)MC
700 700 IF (ABS(Y(K)-C)-E*ABS(Y(K))) 705,705,715
701 701 IF (ABS(YPRM(K)-C)-E*ABS(YPRM(K))) 705,705,715
705 705 IT(K)=I
706      GO TO (500,500,707,709)MC
707 707 C=Y(K)
708      GO TO 755
709 709 C=YPRM(K)
710      GO TO 755
715 715 GO TO (500,500,720,730)MC
720 720 C=Y(K)
725      GO TO 735
730 730 C=YPRM(K)
735 735 CONTINUE

```

Appendix (Continued)

```

740      IT(K)+MAXIT
745      GO TO 755
750 750 C=TIME(K+1)
755 755 DO 770 I=1,NB
760      DO 770 J=1,9
765      N=11-J
770 770 BD(I,N)=(I,N-1)
775      DO 780 I=1,NB
780 780 BD(I,1)=B(I)
785C OUTPUT FOLLOWS
790      PRINT 795
795 795 FORMAT (2X,1HK,4X,4HTIME,7X,4HY(K),8X,5HY'(K),3X,
              10HITERATIONS)
796      IF(M=MAXTM) 802,800
800 800 KSKIP=M/(MAXTM-1)
801      GO TO 805
802 802 KSKIP=1
805 805 DO 810 K=1,M,KSKIP
810 810 PRINT 815,K,TIME(K),Y(K),YPRM(K),IT(K)
815 815 FORMAT (I3,F9.3,2F12.6,I7)
820      GO TO 285
825      END
830      SUBROUTINE COEF(C)
832C THIS SUBROUTINE EVALUATES Z-FORM COEFFICIENTS WHICH MAY
834C BE VARIABLE. FOR TIME-VARYING SYSTEMS THESE
836C COEFFICIENTS ARE FUNCTIONS OF TIME(K). FOR NON-LINEAR
838C SYSTEMS THESE COEFFICIENTS ARE FUNCTIONS OF Y(K)
840C OR Y'(K). "C" REPRESENTS EITHER TIME, Y, OR Y',
842C OR WILL NOT APPEAR AT ALL IN THE A'S AND B'S,
844C DEPENDING ON THE PROBLEM.
848      DIMENSION A(10),B(10)
850      COMMON A,B,T
852C THE A AND B COEFFICIENTS WHICH FOLLOW ARE DIFFERENT
854C FOR EACH PROBLEM.
856      A(1)=C*C+T*T/12.
858      A(2)=C*C+5.*T*/12.
860      B(1)=C*C+T*T/12.
862      B(2)=-2.*C*C+10.*T*T/12.
864      B(3)=C*C+T+T/12.
896      RETURN
898      END

```


VI. Bibliography

- [1] Herbert Freeman, Discrete-Time Systems, New York: John Wiley & Sons, 1965.
- [2] R. Boxer and S. Thaler, "A Simplified Method of Solving Linear and Nonlinear Systems," Proceedings of the IRE, 44, pp. 89-101, January, 1956.
- [3] Henry M. Nodelman and Frederick W. Smith, Mathematics for Electronics with Applications, New York: McGraw-Hill, 1956. pp. 331-337.
- [4] R. Boxer and S. Thaler, "Extensions of numerical transform theory," Rome Air Development Center, Technical Report 56-115, November, 1956.
- [5] Von P. J. Nowacki, "Die behandlung von nichtlinearen problemen in der regelungstechnik (Handling non-linear control problems)," Regelungstechnik, Vol. 8, n. 2, pp. 47-50, February, 1960.
- [6] E. J. Waller, R. R. Reed, Nonlinear Systems--2. Oklahoma State University--Engineering Experiment Station--Publication 134, August, 1963, 101 p.
- [7] Donald R. Coughanowr and Lowell B. Koppel, Process System Analysis and Control. New York: McGraw-Hill, 1965, 384 p.
- [8] K. Hirai, "Analysis of transient response of time-varying control systems by z-transform method," Electrical Engineering in Japan (English translation by Denki Gakkai Zasshi), vol. 85, n. 3, pp. 41-51, March, 1965.
- [9] Maury E. Fowler, "A new numerical method for simulation," Simulation, pp. 324-330, May, 1965.
- [10] A. P. Sage and S. L. Smith, "Real-time digital simulation for systems control," Proceedings of the IEEE, vol. 54, n. 12, pp. 1802-1812, December, 1966.
- [11] H. A. Crosby and D. M. Petersen, "Fortran subroutine solves z-transform inversion," Control Engineering, pp. 92-93, August, 1967.

- [12] K. W. Bach, "Easing z-Transform computations," Electro-Technology, vol. 82, n. 1, p. 48, July, 1968.
- [13] Eliahu I. Jury, Sampled Data Control Systems. New York: John Wiley & Sons, 1958, pp. 290-299
- [14] Rubin Boxer, "A note on numerical transform calculus," Proceedings of the IRE, 45, pp. 1401-1406, October, 1957.
- [15] S. D. Conte, Elementary Numerical Analysis, New York: McGraw-Hill, 1965, pp. 19-23, 108-124.
- [16] W. H. Huggins, "A low pass transformation for z-transforms," Transactions of the IRE, vol. CT-1, pp. 69-70, September, 1954.
- [17] Shepley L. Ross, Differential Equations. New York: Blaisdell Publishing Co., a Division of Ginn and Company, 1964.
- [18] Harold T. Davis, Introduction to Non-linear Differential and Integral Equations. New York: Dover, 1962, pp. 531-537.
- [19] Robert E. Timko, "Analog, digital, and hybrid computer solutions to a nonlinear differential equation," Computer Design, pp. 42-45, September, 1969.
- [20] J. M. Salzer, "Frequency analysis of digital computers operating in real time," Proceedings of the IRE, pp. 457-466, February, 1954.

VII. Vita

Allen Joseph Rushing was born in Charlottesville, Virginia on October 23, 1944. He graduated from Kirkwood High School in Kirkwood, Missouri. Immediately after high school he began studies at the University of Denver in Colorado. In 1966 he received a Bachelor of Science Degree in Electrical Engineering from that university. He then studied journalism and engineering for one semester at the University of Missouri-Columbia.

Since January, 1967, he has worked at the Monsanto Company in St. Louis, Missouri. At Monsanto he began designing electrical facilities for chemical plants. More recently he has developed electronic instruments for process control. Concurrent with employment he has pursued graduate studies in electrical engineering at the St. Louis Graduate Engineering Center of the University of Missouri-Rolla.